Group Project Milestone 1:

Literature Review and Exploratory Data Analysis

Group 5 Members:

Fajar Ibnu Fatihan

Li Jiayi

Lim Ming Wei Jerald

Wang Jinyue

National University of Singapore

IT5006 Fundamentals Of Data Analytics

AY2025/2026 Semester 2

Literature Review

Introduction

A patient's hospital length of stay (LOS) and likelihood of readmission after 30 days are key metrics of healthcare operational efficiency, clinical outcomes and systemic cost (Liu et al., 2024). Prolonged or unnecessary stays increase the risk of adverse events, financial burden, and emotional toll on patients (Graham et al., 2019). These challenges are particularly acute for patients with diabetes, who experience higher hospitalization rates and a propensity for longer, more complex inpatient care (Liu et al., 2024). This review synthesises research on primary factors and machine learning models that predict LOS and 30-day readmission of diabetic patients.

Existing Systems In Literature

Traditional clinical risk scores such as the LACE (LOS, acuity of admission, comorbidity, emergency department visits) score have limited predictive power (Cotter et al. 2012), prompting growing interest in Machine Learning (ML) models that leverage higher dimensionality to improve prediction accuracy (Liu et al., 2024).

Tree-based ensemble methods dominate the literature on LOS and readmission prediction. (Liu et al., 2024; Thenappan et al. 2023) Random Forest (RF) and Extreme Gradient Boosting (XGBoost) consistently emerge as top performers due to their ability to capture nonlinear interactions and high-order feature relationships common in clinical datasets (Liu et al., 2024). Liu et al. (2024) reported F1 scores of 0.83 and 0.84 for RF and XGBoost respectively in 30-day readmission prediction. In predicting LOS, the RF model also achieved an R^2 score of 0.856 and a Mean Absolute Error (MAE) lower than 0.45 (Morton et al., 2014; Naila et al., 2020). The R^2 score indicates that the model could explain about 85.6% of the variance in the length of stay. RF is valued for robustness and interpretability, whereas XGBoost is more scalable, incorporates regularization to reduce overfitting, and handles sparse or missing values effectively.

Support Vector Machines (SVMs) have also been applied to readmission prediction, achieving an F1 score of 0.82 in Cui et al.'s (2018) study. While less interpretable than tree-based models, SVMs excel in high-dimensional spaces and are less sensitive to outliers. (Cui et al., 2018)

Modeling Approaches

Various strategies were chosen when partitioning datasets for training and testing. Shang et al. (2021) used an 80:20 train-test split while others employed k-fold cross validation, a more robust approach that reduces variance in model evaluation. Typically, a 5-fold cross validation method was used. (Cui et al., 2018; Liu et al., 2024) The source dataset was divided into five parts, with each of these five parts taking turns as test data, while the remaining four parts are used as training data. (Cui et al., 2018)

To address class imbalance, three resampling strategies are often used: over-sampling, under-sampling or a hybrid of the two. (Bach & Aarseth, 2016) To address the class imbalance where non-readmitted cases greatly outnumbered readmitted ones, many studies employed synthetic minority oversampling technique (SMOTE). SMOTE uses the nearest neighbours of the minority class samples to

generate synthetic samples of the minority class to improve model sensitivity to readmissions (Cui et al., 2018).

Ensemble models combine multiple learners to boost predictive accuracy. In RF, bootstrap sampling generates diverse training subsets, and each decision tree is built using a random subset of features, reducing correlation between trees and improving generalization (Thenappan et al. 2023). XGBoost extends gradient boosting with shrinkage, regularization, and efficient handling of missing values, enabling high predictive accuracy while mitigating overfitting. In overfitting mitigation, after each new tree is added, its contribution to the final prediction is scaled down (Chen & Guestrin, 2016). This prevents any single tree from having too much influence. There is also a penalty imposed on model complexity. If a tree becomes too complex, it gets penalized. This forces the model to remain simpler. In handling missing values, XGBoost has a built-in ability to learn the best way to handle missing values on its own during the training process, which saves time and can lead to better predictions.

Feature Engineering Techniques

Feature engineering is central to model performance. Standard preprocessing steps include removal of duplicates, dropping columns with a high percentage of missing values, and dropping columns with only one unique value. It was also common practice to use Z-score normalisation (Liu et al., 2024; Shang et al., 2021) over min-max standardisation of numerical data (Cui et al., 2018).

One advanced technique for handling missing data was Multiple Imputation with Chained Equations (MICE). MICE was shown to better preserve the data's underlying structure compared to simple mean/median imputation. For the high-cardinality diagnostic codes, an effective strategy is to leverage the hierarchical nature of the coding system by grouping codes into broader clinical categories, which reduces dimensionality while retaining clinical meaning (Ayden et al., 2024).

Optimisation of the number of features also played a role in improving the predictive accuracy of the algorithms while making models less computationally expensive. One study utilised the Greywolf optimiser (GWO) algorithm, which helped reduce the number of features while retaining the number of encounters in the dataset. (Liu et al., 2024) GWO mimics the hunting behavior of grey wolves to balance exploration and exploitation, converging on feature subsets with high predictive fitness. (Mirjalili et al., 2014)

Evaluation Metrics of ML Models

Studies employed a mix of threshold-dependent and threshold independent metrics to evaluate the performance of each ML model. Threshold dependent metrics included accuracy, precision, recall and F1 score, with F1 often preferred due to its balance of false positives and false negatives in imbalance datasets. Threshold-independent metrics such as the area under the receiver operating characteristic curve (AUROC) scores reflect the models' performance across various threshold settings (Liu et al., 2024). Additional metrics were sometimes applied. Cui et al. (2018) used the geometric mean (G-mean) of class recall rates to evaluate the performance on imbalance datasets.

Conclusion

This review highlights a clear trend in the field: a move away from traditional scoring systems toward more sophisticated machine learning models for predicting diabetic patient outcomes. Key Trends show that tree-based ensemble methods, particularly Random Forest and XGBoost, are consistently the top performers due to their high accuracy. However, there are notable gaps in the current literature. While the predictive power of these models is well-established, there is less focus on model interpretability. It is often unclear why a model makes a specific prediction, which is a barrier to clinical trust and adoption. One area of improvement might be the skew of racial data towards Caucasians over other races, which might make these models less robust in predicting readmission in other racial groups.

References

1. Liu, V., Sue, L., & Wu, Y. (2024). Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. Journal Of Medical Artificial Intelligence, 7. doi:10.21037/jmai-24-70
2. Graham, E., Saxena, A., & Kirby, H. (2019). Identifying high risk patients for hospital readmission. SMU Data Science Review, 2(1), 22.
3. Cotter, P. E., Bhalla, V. K., Wallis, S. J., & Biram, R. W. (2012). Predicting readmissions: poor performance of the LACE index in an older UK population. Age and ageing, 41(6), 784–789. https://doi.org/10.1093/ageing/afs073
4. Thenappan, S., Ramshankar, N., Hemavathy, N., Subashree, V., Manjul, R. R., & Rajeswari, J. (2023, July). Machine Learning Classifiers to Decrease Diabetic Patients Probability of Hospital Readmission. In 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 829-834). IEEE.
5. Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., & Kakadiaris, I. A. (2014, December). A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In 2014 13th International Conference on Machine Learning and Applications (pp. 428-431). IEEE.
6. Naila, Mekhaldi & Caulier, Patrice & Chaabane, Sondes & Chraibi, Abdelahad & Piechowiak, Sylvain. (2020). Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting. 10.1007/978-3-030-45688-7_21.
7. Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. Computer methods and programs in biomedicine, 166, 123–135. https://doi.org/10.1016/j.cmpb.2018.10.012
8. Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. BMC medical informatics and decision making, 21(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y
9. Bach, A. S., & Aarseth, H. (2016). Adaptation, equality, and fairness. Towards a sociological understanding of ‘the supportive husband’. Norma, 11(3), 174-189.
10. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.
11. Ayden, M. A., Yuksel, M. E., & Yuksel Erdem, S. E. (2024). A two-stream deep model for automated ICD-9 code prediction in an intensive care unit. Heliyon, 10(4), e25960. https://doi.org/10.1016/j.heliyon.2024.e25960
12. Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

Exploratory Data Analysis Report

Dataset Overview and Data Quality (Figure 1)

The dataset contains 101,766 patient records with 50 variables capturing demographic information, clinical indicators, and treatment outcomes. A data quality assessment was conducted to evaluate each variable's missing data ratio. Three key variables, weight (96.9% missing), max glucose serum (94.7% missing) and A1C results (83.3% missing), have the highest missing ratio and are not useful for our analysis. Therefore, we excluded them from the subsequent analysis.

Age and Gender Demographics (Figure 2)

Analysis of patient demographics reveals a strong age-related pattern in diabetes prevalence. The age distribution shows a clear upward trend, with the number of patients increasing significantly after age 40 and peaking between 60-80 years old. This pattern aligns with known epidemiological trends where diabetes risk escalates with advancing age. However, gender analysis shows roughly equal distribution between male and female patients, with both genders following similar age-related trends. This suggests that while age is a significant risk factor for diabetes diagnosis, gender plays a minimal role within this hospital population.

Racial Distribution (Figure 3)

The dataset shows significant racial skew, with Caucasian patients representing approximately 76% of the population, followed by African American patients at roughly 20% . Hispanic, Asian, and other racial groups each comprise less than 5% of the dataset. This distribution reflects the higher prevalence of diabetes in Caucasian and African American populations.

30-Day Readmission Patterns by Demographics (Figure 4)

Surprisingly, patients in the 20–30 age group show the highest 30-day readmission rate (14.24%), even higher than older, more medically complex groups. This counterintuitive result suggests that risk in younger adults may be shaped less by clinical complexity and more by behavioral elements, such as inconsistent treatment adherence or lifestyle patterns. While the dataset does not allow us to confirm these mechanisms directly, the finding points to an area that warrants further exploration.

Healthcare Utilization and Readmission Risk (Figure 5)

Strong positive correlation is discovered between prior healthcare utilization and 30-day readmission risk. Patients with higher numbers of emergency department visits and previous inpatient admissions show substantially elevated readmission rates. These patterns suggest that historical healthcare utilization serves as a powerful predictor of future readmission risk.

Medical Complexity as a Readmission Predictor (Figure 6)

A new feature, medical complexity score, calculated as the average of number of medications, number of procedures, and number of laboratory tests, is interpreted as a strong predictor of 30-day readmissions. Readmission rates climbing dramatically as complexity increases. Especially, patients with the complexity scores above 40 have a clear upward trend of 30-day readmission rate. This finding suggests that identifying and intensively managing the small subset of highly complex patients could be an effective strategy for reducing overall readmission rates.

Length of Stay Analysis (Figure 7)

Hospital length of stay (LOS) exhibits a clear distribution pattern, with 48% of patients having short stays (1-3 days), 36% having medium stays (4-7 days), and 16% requiring long stays (8+ days). We can see a clear trend from the figure that Age has a strong relationship with LOS, showing a dramatic shift from predominantly short stays in younger patients to increasingly longer stays in older age groups.

Key Findings and Implications

This analysis identifies medical complexity as a powerful predictor of 30-day readmissions, surpassing traditional demographic factors. This suggests targeted interventions for high-complexity patients could yield significant improvements. The counterintuitive finding of highest readmission rates in young adults warrants further investigation into age-specific care protocols. Healthcare utilization patterns, particularly emergency department usage and prior admissions, serve as strong early warning indicators for readmission risk.

Conclusion

Healthcare systems should implement risk stratification protocols using medical complexity scores and prior utilization patterns to identify high-risk patients early in their hospital stay. Specialized care coordination teams should focus on patients with multiple comorbidities, extensive medication regimens, and histories of frequent healthcare utilization. Age-specific discharge planning protocols may be necessary, particularly for young adult patients who show unexpectedly high readmission rates. Length of stay should be considered as both an outcome measure and a risk factor, with extended stays triggering enhanced discharge planning and post-acute care coordination efforts.

Dashboard link:
https://public.tableau.com/views/DiabeticPatientTreatmentHistory/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

Appendix

	missing_ratio
weight	0.968585
max_glu_serum	0.947468
A1Cresult	0.832773
medical_specialty	0.490822
payer_code	0.395574
race	0.022336
diag_3	0.013983
diag_2	0.003518
diag_1	0.000206
patient_nbr	0.000000
time_in_hospital	0.000000
admission_source_id	0.000000
num_lab_procedures	0.000000
encounter_id	0.000000
admission_type_id	0.000000

Figure 1

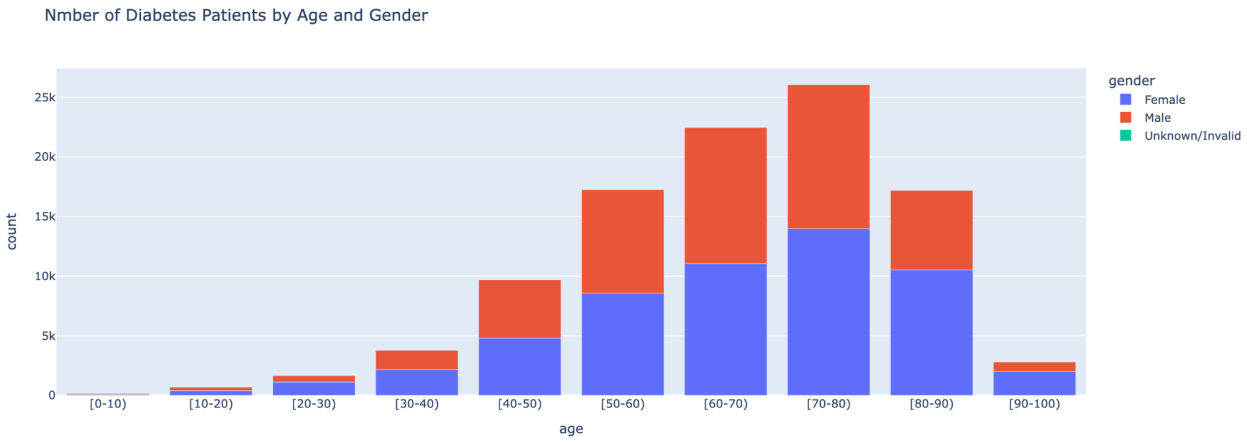

Figure 2

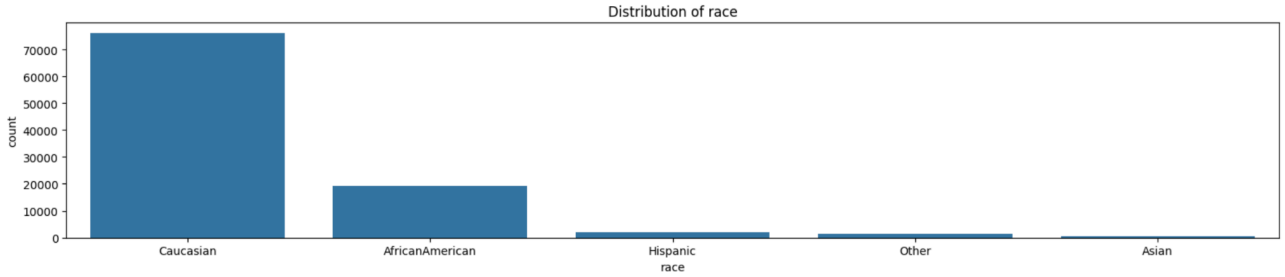

Figure 3

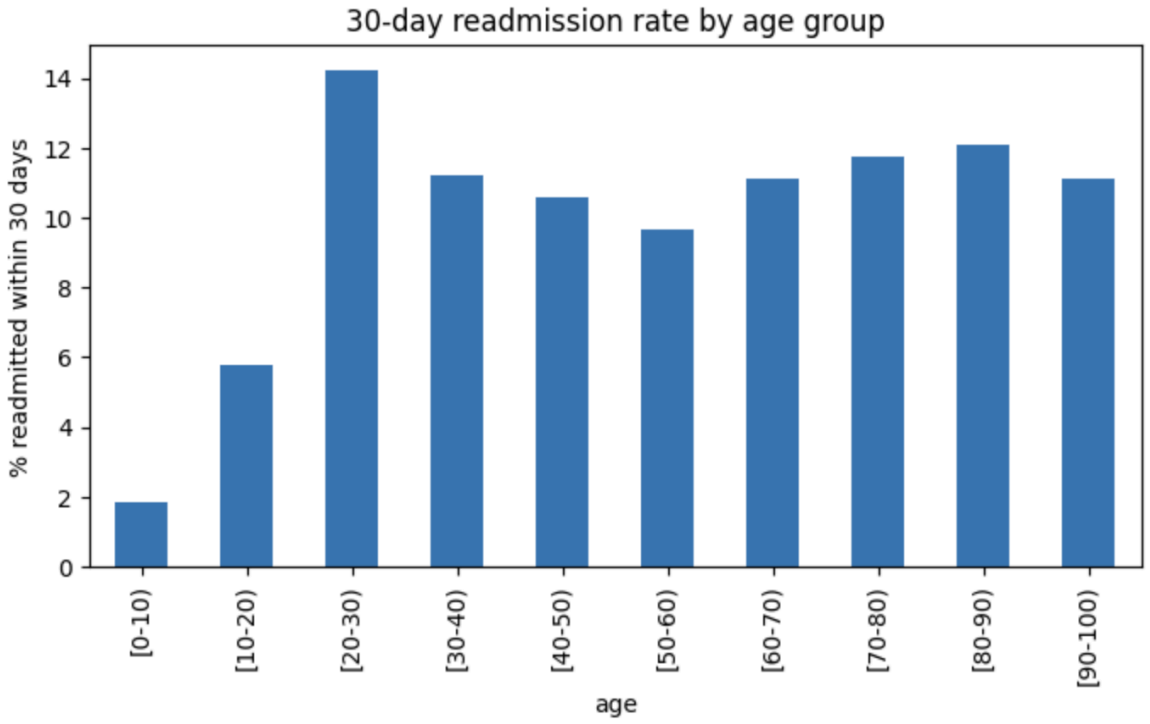

Figure 4

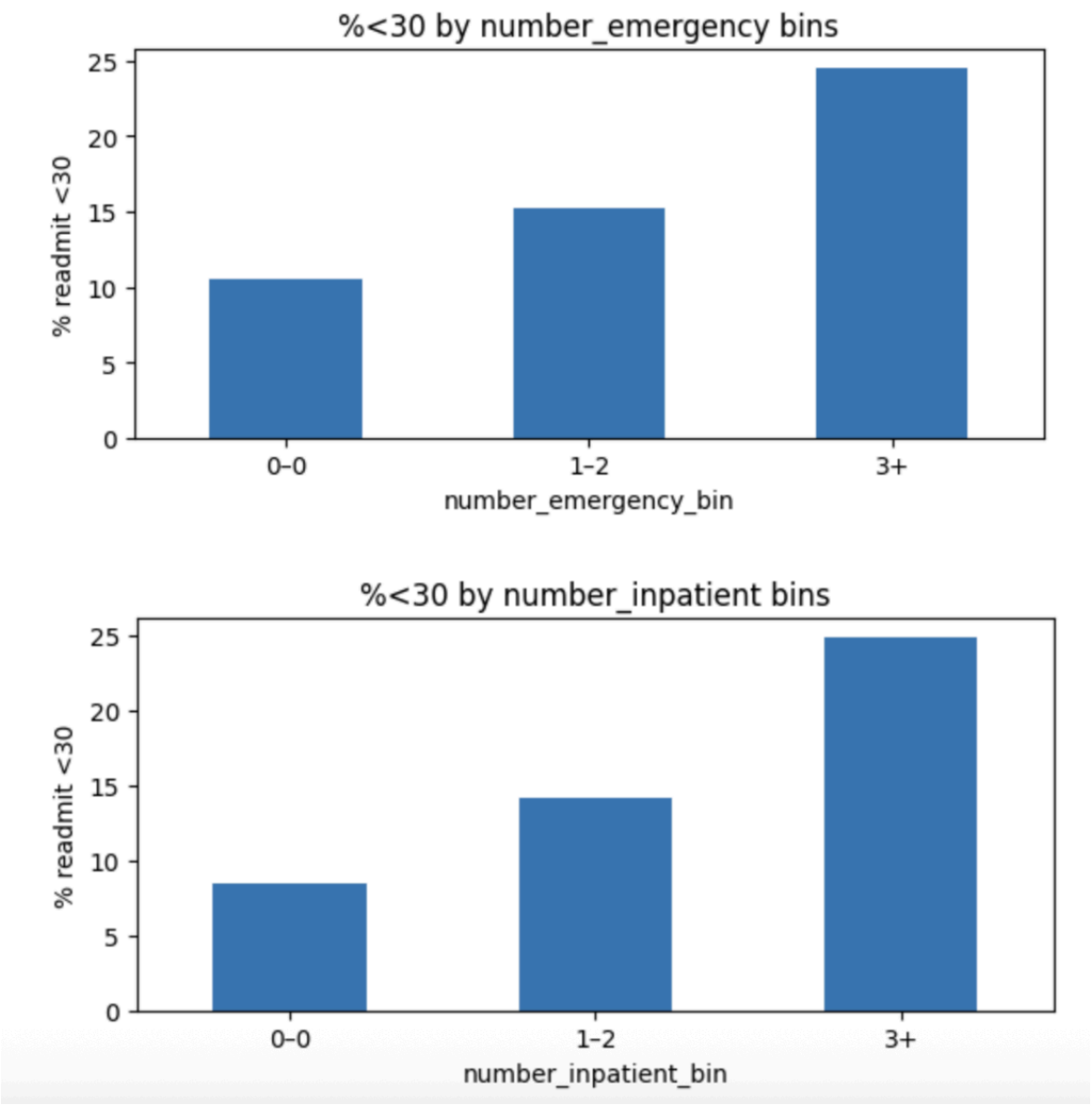
%<30 by number_emergency bins
% readmit <30
25
20
15
10
5
0
0–0
1–2
3+
number_emergency_bin
%<30 by number_inpatient bins
% readmit <30
25
20
15
10
5
0
0–0
1–2
3+
number_inpatient_bin

Figure 5

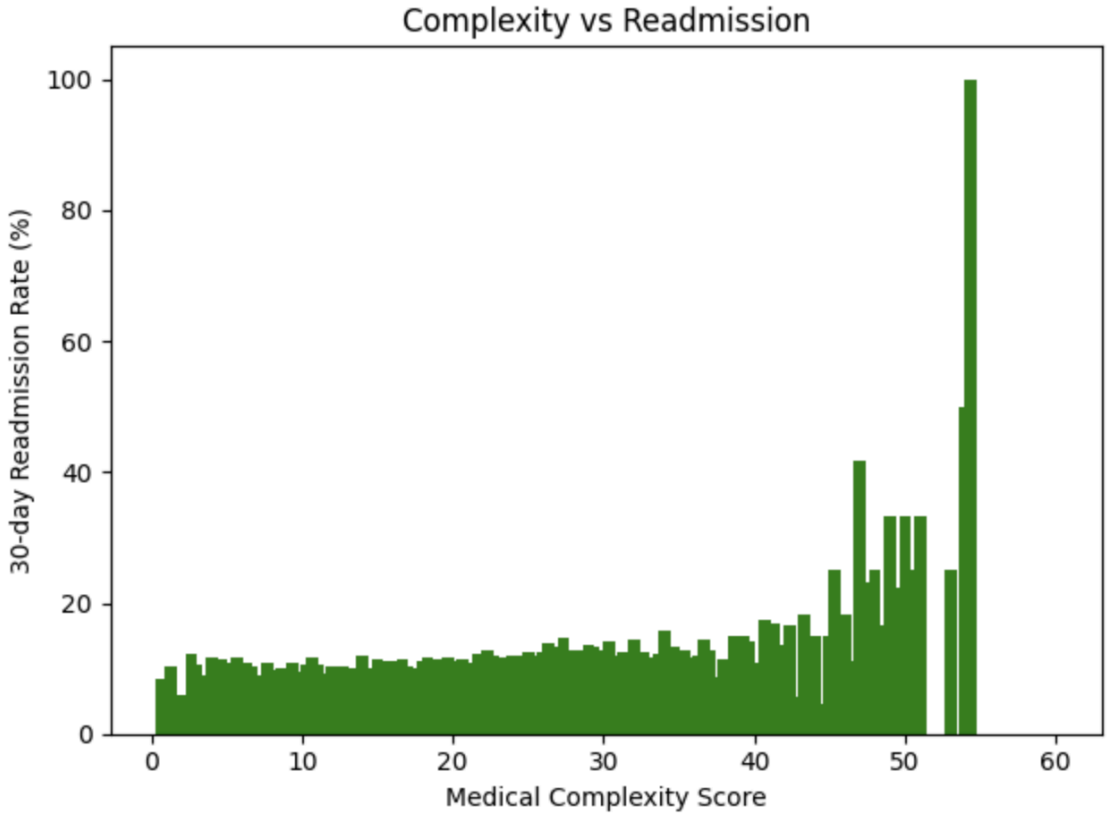

Figure 6

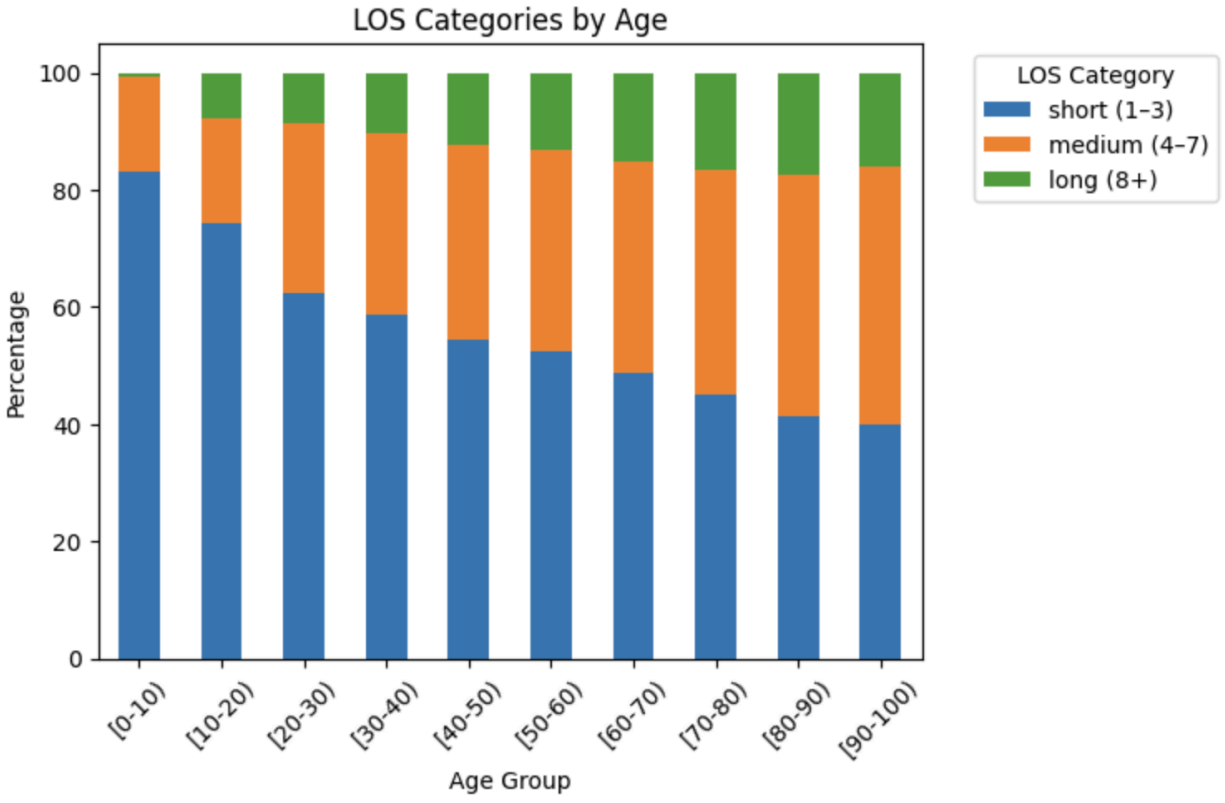

Figure 7